Journal of Modern Applied Statistical Methods

Volume 7 | Issue 2 Article 3

11-1-2008

Comparing Factor Loadings in Exploratory Factor Analysis: A New Randomization Test

W. Holmes Finch
Ball State University, whfinch@bsu.edu

Brian F. French
Washington State University, frenchb@wsu.edu

Journal of Modern Applied Statistical Methods
November, 2008, Vol. 7, No. 2, 376-384

1538 – 9472/08/$95.00

REGULAR ARTICLES

Comparing Factor Loadings in Exploratory Factor Analysis: A New Randomization Test

W. Holmes Finch
Ball State University

Brian F. French
Purdue University

Factorial invariance testing requires a referent loading to be constrained equal across groups. This study introduces a randomization test for comparing group exploratory factor analysis loadings so as to identify an invariant referent. Results show that it maintains the Type I error rate while providing adequate power under most conditions.

Key words: Exploratory factor analysis, randomization test, multigroup confirmatory factor analysis, invariance testing.

Introduction

Score validity evidence can be considered the primary focus in instrument development and evaluation (AERA, APA, & NCME, 1999). For instance, Standard 1.1 of the *Standards for educational and psychological testing* states "A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation" (p. 17, AERA et al., 1999). Measurement invariance (MI) or equivalence is one form of validity evidence that is important when scores are used for group comparisons. MI refers to the case where an assessment measures one or more latent constructs identically across groups. The presence of this property helps ensure that the measurement of the specified construct is the same across groups, thus allowing for accurate comparisons in score parameters. Otherwise group comparisons may be meaningless, as observed differences could be the result of ability differences or measurement differences.

Factor invariance is one form of measurement invariance (MI) and is typically established using multi-group confirmatory factor analysis (MCFA). Through MCFA, an *a priori* theoretically specified latent structure of an instrument is evaluated for MI across groups (Alwin & Jackson, 1981; Golembiewski, Billingsley, & Yeager, 1976). The presence of MI is tested using differences in the chi-square goodness-of-fit statistics for more (loadings held equal across groups) and less restrictive (loadings allowed to vary by group) models. If the fit of the models differs significantly, as measured by the chi-square difference test, the researcher concludes a lack of invariance. This method is well documented (e.g., Bollen, 1989; Byrne, Shavelson, & Muthén, 1989; Jöreskog & Sörbom, 1996; Maller & French, 2004; Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993).

The requirement of an equality constraint of a referent indicator across groups in MCFA calls for methodological attention (Millsap, 2005). Comparison of a latent factor model can only occur if the same coordinate system is used for all groups in question (Wilson, 1981). Model identification procedures

W. Holmes Finch is Professor of Psychology in the Department of Educational Psychology, and Educational Psychology Director of Research in the Office of Charter School. Email: whfinch@bsu.edu. Brian F. French is Associate Professor and Co-Director Learning and Performance Research Center Washington State University. Email: frenchb@wsu.edu

ensure this required comparability by assigning the same units of measurement to the latent variables for groups in question (Jöreskog & Sörbom, 1996). Model identification is often accomplished by assigning the latent factors to a scale based on a common indicator across groups, typically either a factor variance or a factor loading for a single variable. The most common practice is to set one of these parameter values to 1.0 across groups, with the factor loading method being the most common (Brown, 2006; Vandenberg & Lance, 2000). This factor loading referent approach requires the assumption that the referent loading is equal for all groups in the population (i.e. the loading is assumed to be invariant).

When the referent parameter is not invariant, estimates of other model parameters, such as factor loadings, may be distorted and hypothesis tests for the group invariance of these other parameters may be inaccurate (Bollen, 1989; Cheung & Rensvold, 1999; Millsap, 2001). Therefore, a circular situation exists where (a) the referent loading must be invariant, (b) invariance of the referent (or any other) loading cannot be established without estimating a model, and (c) model estimation requires an invariant referent loading. Thus, we return to the original invariant referent assumption, which is commonly not assessed in practice, most likely due to the fact that there is not a relatively straight forward way of doing so. A procedure to locate an invariant referent variable would be useful to ensure the remainder of invariance assessment is accurate.

Heretofore, this assumption of referent invariance could not be directly tested (Bielby, 1986; Cheung & Rensvold, 1999; Wilson, 1981). A search procedure, the factor-ratio test and stepwise partitioning procedure, has been suggested (Rensvold & Cheung, 2001). The procedure uses each variable as the referent in a set of models with each other variable constrained to be invariant. The iterative procedure tests all pairs of variables (i.e., p (p – 1) / 2 pairs) and becomes quite complex as the number of indicator variables increases, making it not "user-friendly" for practitioners (Vandenberg, 2002). For example, a moderate length instrument (i.e., 30 indicators) requires 435 individual invariance tests to fully identify which loadings could be used as a referent in the final MCFA analysis. Evaluation of this procedure demonstrated adequate (e.g., acceptable false and true positives) but not perfect performance (French & Finch, 2006a).

Exploratory factor analysis (EFA) has been suggested as an alternative approach for identifying an invariant referent loading. In its relative simplicity, EFA overcomes the limitations associated with the factor-ratio test and search procedure. The EFA based approach involves conducting a single EFA for each group separately and descriptively comparing their respective loading estimates to ascertain which appear to be invariant in the sample. Such an analysis may be considered a weak test of factorial invariance (Zumbo, 2003) and is in accord with suggestions that EFA be used to examine loadings with an "interocular eyeball test" (Vandenberg, 2002, p. 152) to judge the similarity of loadings to identify referent variables. Evaluation of this procedure has been favorable (Finch & French, *in press*), though it does not offer a formal hypothesis test of invariance, instead allowing for the comparison of parameter estimates across groups in order to provide a sense of factor loading differences without the need to conduct a large number of analyses. Specifically, pattern coefficients appearing most similar would be eligible for serving as a referent variable in the MCFA. The obvious limitation to the current EFA procedure is the lack of a statistical test to give a formal determination about the differences between factor loadings.

The purpose of this study was to develop a randomization test based on EFA and to assess its utility in identifying invariant factor loadings between two groups. This procedure would be used prior to conducting the actual MCFA, as a purification process for identifying a loading that is likely to be group invariant and thus eligible for use as the referent parameter. The procedure entails conducting one EFA per group and then comparing the factor loadings (i.e., pattern coefficients) from the separate analyses via the test statistic to determine differences of individual loadings. Loadings that are significantly different would not be used as a referent.

Factor loading invariance randomization test (FLIRT)

Statisticians have developed exact tests for a number of applications involving group comparisons (see Good, 1994, for a thorough description of exact tests). Regardless of the context, every exact test for group comparison involves finding all possible permutations of the data, with respect to group membership. For each of these permutations the test statistic of interest is calculated and the collection of these statistics across all permutations forms a sampling distribution. The test statistic for the observed sample is also calculated and, if it is more extreme than a predetermined (e.g., 95th) percentile of the permutation distribution, the null hypothesis of no group difference can be rejected.

One common problem in the actual application of permutation tests is that, even for modestly sized samples, the number of permutations that must be determined can be large. For example, for a simple two group comparison with a total sample of 30 individuals (15 per group), the number of permutations would be 155,117,520. The computer time necessary to conduct analyses for each of these permutations would be prohibitive for any real application. An alternative approach to using all possible permutations is known as randomization, or Monte Carlo, testing (Edgington, 1980). With this methodology, a random sample of the permutations is selected and the test statistic of interest is calculated for each to create the sampling distribution as described above. As with the full permutation testing approach, the test statistic value obtained from the observed data is compared with this distribution and, if it is more extreme than some predetermined (e.g. 95th) percentile, the null hypothesis of no group difference is rejected. The description of the specific randomization test statistic for comparing two groups' factor loadings appears below.

The factor loading invariance randomization test (FLIRT) for comparing two groups' factor loadings is based upon the supposition that there exists configural invariance for the two groups; i.e., the basic factor structure is the same, though the actual factor loading values may not be. To test the null hypothesis of equal (invariant) group loadings for a single indicator variable, EFA is run separately for the two groups and the difference in the loadings for the target indicator is calculated. Next, 100 random samples are taken from the population of all possible permutations and for each of these EFA is conducted by group. The difference in the target loadings is calculated for each permutation to develop a distribution against which the group loading difference for the observed data is compared. If this observed difference is larger than the 95th percentile from the randomization distribution, the null hypothesis of no group differences on the target loading is rejected. The current study evaluated FLIRT through the use of a Monte Carlo simulation, as well as the analysis of a real dataset. The performance of the test was judged in terms of power and Type I error under a variety of conditions (e.g., sample size, factor model) in the simulation study, and by comparing hypothesis test results for the observed data with those presented in Thompson (2004).

Methodology

Simulated data were used to control variables that could influence the magnitude of factor loading estimates, with 1,000 replications for each combination of conditions described below. Simulations and analyses were completed in *SAS, V9.1* (The SAS Institute, 2003). Conditions were held as consistent as possible with previous studies (e.g., Finch & French, 2008 *in press*) for comparability of results. Second, a real data set, the LibQUAL+ study (Thompson, 2004), was employed to provide an applied example.

Number of Factors and Indicators

Data were simulated from both 1- and 2-factor models, with interfactor correlations set at .50 to represent moderately related factors, and simple structure for continuous and normally distributed subtest level data. The number of indicators per factor was 6.

Sample Size

The necessary sample size to obtain reasonable estimates in factor analysis varies

depending on the data conditions. Four sample size conditions were simulated: 100, 250, 500, and 1,000 per group in order to reflect small, medium and large samples. These values are consistent with other factor analysis simulation studies (Cheung & Rensvold, 2002; Lubke & Muthén, 2004; Meade & Lautenschlager, 2004), ranging from poor (n = 100) to excellent (n = 1,000) (Comery & Lee, 1992), and may not be of much concern here as communalities were high (MacCallum, Widaman, Zhang, & Hong, 1999).

Magnitude of Difference with the Non-Invariant Indicators

Six levels of factor loading values for the non-invariant indicator were simulated. A baseline condition was established where no group differences in loadings were present, with all variables having a loading value of 0.75, including the target. The remaining 5 conditions were characterized by declines in the target loading from 0.10 to 0.50 in increments of 0.10 (i.e., 0.65, 0.55, 0.45, 0.35, and 0.25). This wide range of levels was selected since there is no effect size, at least to our knowledge, for what represents a meaningful difference (Millsap, 2005) and the range covers previously used values in MCFA simulation work (e.g., French & Finch, 2006b; Meade & Lautenschlager, 2004).

Contamination

The location of invariant parameters may be influenced by the number of indicators that lack invariance (Millsap, 2005; Yoon & Millsap, 2007). Thus, the presence of a factor loading, other than for the target indicator, exhibiting a group difference was varied as either present or absent. In other words, for half of the simulated conditions only the target indicator loading was contaminated, while for the other half of the simulations a second target indicator loading also was contaminated at the same difference as the target indicator. This allowed assessment of the influence of additional contaminated variables.

Analysis

All analyses were conducted by group using maximum likelihood factoring with PROMAX rotation in the 2-factor condition. These settings follow recommendations for using EFA for a referent indicator search and are more consistent with educational and psychological data (e.g., presence of measurement error, correlated factors; (Vandenberg, 2002).

Evaluation Criteria

The outcomes of interest for this study were the power and Type I error rates of the FLIRT. Specifically, the Type I error rate was calculated as the proportion of simulation replications for which the test statistic rejected the null hypothesis when the groups' loadings on a target indicator did not differ. In similar fashion, power was calculated as the proportion of the simulation replications for which the test statistic rejected the null hypothesis when the groups' loadings on the target indicator did in fact differ. To determine which conditions influenced the outcomes of interest, ANOVA and variance components analysis were used with each of the manipulated factors serving as an independent variable. For the applied data set results are presented in terms of locating differences in factor loadings as would be for an application.

Results

Simulation study

Type I error

None of the manipulated factors, or their interactions, was identified by the ANOVA as being significantly related to the Type I error rate of the FLIRT. Table 1 contains these Type I error rates by each of the manipulated variables. Overall, there is a very slight elevation of the error rate above the nominal 0.05, with the most notable difference between the 1 and 2 factor conditions. However, none of the sample differences evident in this table were statistically significant, suggesting that they may not be present in the population as a whole.

Power

Based on the results of the ANOVA and variance components analysis, the interaction of sample size by the difference in the groups' target loadings, as well as the main effects of

Table 1: Type I Error Rates by Sample Size, Number of Factors, and Level of Contamination

Sample size	Type I error rate
100	0.067
250	0.064
500	0.059
1000	0.060
Factors	
1	0.069
2	0.057
Contamination	
No	0.061
Yes	0.064

Table 2: Power by Sample Size and Group Difference in Target Loading

Sample size per group	Difference	Power
100	0.1	0.23
	0.2	0.61
	0.3	0.87
	0.4	0.96
	0.5	0.97
250	0.1	0.49
	0.2	0.92
	0.3	0.96
	0.4	1.00
	0.5	1.00
500	0.1	0.80
	0.2	1.00
	0.3	1.00
	0.4	1.00
	0.5	1.00
1000	0.1	0.97
	0.2	1.00
	0.3	1.00
	0.4	1.00
	0.5	1.00

sample size and difference in target loadings were statistically significant and contributed more than 10% of the variance to the power of the test statistic. Specifically, the interaction accounted for 38.4% of the variance as did the main effect of difference in loading values, while the main effect of sample size contributed an additional 20.2% to the variation of power. contains power rates by the interaction of sample size and group loading differences.

For the largest sample size condition, power was well above 0.95 regardless of the difference between the groups' loadings. Thus, even when the target loadings only differed by 0.1 the test statistic would virtually always identify this divergence. On the other hand, for samples of 100 per group, the test had power rates below 0.8 for differences of 0.1 and 0.2. In general, across the lower sample size conditions (100 and 250 most particularly), power was

somewhat low for a difference of 0.1 but rose to above 0.8 for discrepancies in target loadings of 0.3 or more.

Table 3 shows power rates by the number of factors and level of contamination. Neither of these terms contributed more than 3% to the variance in power. A perusal of the results in this table shows that there were essentially no differences in power for 1 and 2 factors or when another loading beyond the target loading differed between the groups.

Table 3: Power by Number of Factors and Contamination

Number of factors	Power
1	0.90
2	0.88
Contamination	
No	0.89
Yes	0.89

Analysis of real data

To demonstrate the FLIRT in real world conditions, data taken from the LibQUAL+ study were analyzed. For a more complete discussion of this dataset and the study from which it was drawn, the interested reader is encouraged to consult Thompson (2004). The 12 items included on this survey could be divided into three factors, including service provided by library staff, the environment of the library and the quality of the library's holdings. Each factor was represented by 4 items, which were on a rating scale with response options ranging from 1 to 9. The dataset used, which is available in Thompson (2004), included a random sample of 200 survey respondents, 100 of whom were graduate students and 100 who were faculty members.

Thompson described differences in factor loading values between graduate students and faculty members for item 6, "A meditative place". To demonstrate the utility of the FLIRT with real data, the faculty and student loadings for item 6 were compared using this new statistic. The factor loading values by group were 0.7587 for graduate students and 0.9079 for faculty, leading to an observed loading difference of 0.1492. The distribution of differences across the 100 randomized datasets appears in Figure 1, a visual examination of which shows that the observed difference falls in the 99th percentile of the randomized values. Thus, if α= 0.05, we would conclude that there is a statistically significant difference between the loading values for the two groups, which is in line with the conclusion reached by Thompson. The two groups loadings for item 5, "A haven for quiet and solitude", were also compared. This was not identified by Thompson as differing between the groups. The loading for the students was 0.9114, and 0.9342 for the faculty, leading to an observed difference of 0.0228. This value fell at the 46th percentile of the randomized loading differences, which would lead to a conclusion of no significant difference between group loadings at the aforementioned level of 0.05.

The purpose of this analysis with previously analyzed real data using MCFA was to demonstrate the potential utility of FLIRT. If FLIRT had been used as a step prior to the MCFA in this example, item 6 would not have been selected as a referent variable whereas item 5 could have been. The results presented are in accord with those of Thompson (2004), thus providing further evidence, beyond the simulation study, that this new statistic does appear to be reasonably accurate in correctly identifying group loading differences, even for samples as small as 100 per group.

Conclusion

The results suggest that in many instances, the FLIRT may be a useful tool for identifying potential indicator variables with invariant factor loadings across groups for use in a subsequent MCFA. This outcome was especially evident when the differences between loadings and/or the sample sizes were large. However, even for differences in loadings as small as 0.2 and samples of 100 per group, FLIRT was able to find differences more than 60% of the time. In all but one case, when sample size was 250 or more per group, the rates for correctly detecting loading differences were at least 0.8, and often near 1.0. Furthermore, the Type I error rates (identifying loadings as differing when they do not) were very close to the nominal rate of 0.05

for all studied conditions. The combination of these results supports the use of the new FLIRT statistic in conjunction with EFA for accurately detecting a non-invariant loading that could then be used as the referent in a subsequent MCFA.

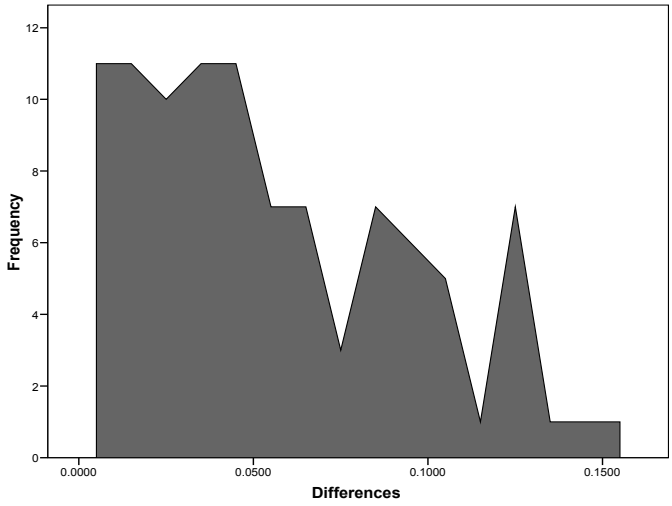
Correct specification of an invariant referent loading is a crucial step in MCFA. Failure to do so could lead to biased parameter estimates and, in turn, compromise other analyses, such as latent mean comparisons. The primary method suggested for identifying invariant indicators is the factor-ratio test and SP procedure (Rensvold & Cheung, 2001), which can be a very complex and time consuming multi-step technique. While this procedure does work reasonably well in identifying invariant referent loadings, it can become intractably time consuming with increasing model complexity (French & Finch, 2006a). To overcome such limitations, EFA is one approach that has been advocated for use in practice and involves comparison of factor loading estimates between two groups (Vandenberg, 2001; Zumbo, 2003). While this method does not have the advantage of significance testing that is offered by the factor-ratio test, it is much simpler to conduct. We have attempted to overcome the inference limitation of EFA, while maintaining its advantage of simplicity, by developing the FLIRT.

The results seem to indicate that in need to locate an invariant referent for use in MCFA they may find that this simple approach performs well in a fairly wide variety of study FLIRT generally provides an accurate conditions such as those simulated; EFA with assessment of identifying the variables that may lack invariance. Therefore, when practitioners conditions. The FLIRT is more accurate (i.e., greater power) with larger sample sizes and a greater magnitude of difference between loadings and appears to have Type I error rates that are always close to the nominal level.

Limitations and directions for future research

The generalizability of the results is limited to the conditions simulated in this study. First, the factor models examined were fairly simple (1 or 2 factors with 6 indicators each). Thus, in future research the FLIRT should be evaluated with more complex models and data (e.g., greater number of factors, different variables, various levels of communalities). Second, a related area that deserves attention is the combination of loadings for the observed variables. In this study, all of the loadings were set at 0.75 (unless contaminated). Given that this is the first investigation of the randomization test to accurately identify invariant referent variables, clarity of result interpretation was considered paramount, and thus non-target loadings were not varied. However, further investigation should be carried out for a more complex combination of loading values and factor models, as well as data conditions (e.g., ordinal variables) before the test is applied unequivocally.

Figure 1: Distribution of randomized loading differences for item 6



References

Finch, W. H., & French, B. F. (2008). Using exploratory factor analysis for locating invariant referents in factor invariance studies. *Journal of Modern and Applied Statistical Methods, 7(1), 223-233*.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Alwin, D. F. & Jackson, D. J. (1981). Applications of simultaneous factor analysis to issues of factorial invariance. In D. Jackson & E. Borgatta (Eds.), *Factor analysis and measurement in sociological research: A multi-dimensional perspective* (pp. 249-279). Beverly Hills, CA.: Sage.

Bielby, W. T. (1986). Arbitrary metrics in multiple-indicator models of latent variables. *Sociological Methods & Research*, *15*, 3-23.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York, NY: Wiley.

Brown, T.A. (2006). *Confirmatory Factor Analysis for Applied Research*. The New York, NY: Guilford Press.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures. *Psychological Bulletin*, *105*, 456-466.

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*, 1-27.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis.* Hillsdale, N.J.: L.Erlbuam Associates.

Erlbaum. Edgington, E.S. (1980). *Randomization Tests.* New York, NY: Marcel and Dekker, Inc.

French, B. F., & Finch, W. H. (2006a, June). *Locating the Invariant Referent in Multi-Group Confirmatory Factor Analysis.* Paper presented at the International Psychometric Society meeting in Montreal, Canada.

French, B. F., & Finch, W. (2006b). Confirmatory factor analytic procedures for the determination of measurement invariance. *Structural Equation Modeling*, *13*, 378-402.

Golembiewski, R.T., Billingsley, K. & Yeager, S. (1976). Measuring change and persistence in human affairs: Types of change generated by OD designs. *Journal of Applied Behavioral Science, 12,* 133-157.

Good, P. (1994). *Permutation Tests: A practical guide to resampling methods for testing hypotheses.* New York, NY: Springer-Verlag.

Jöreskog, K.G., & Sörbom, D. (1996). *LISREL 8: User's reference guide.* Chicago: Scientific Software.

Lubke, G. H., & Muthén, B. O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11*, 514-534.

MacCallum, R.C., Widaman, K.F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4,* 84-99.

Maller, S. J., & French, B. F. (2004). Factor invariance of the UNIT across deaf and standardization samples. *Educational and Psychological Measurement*, *64*, 647-660.

Meade, A. W., & Lautenschlager, G. J. (2004). A monte-carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling*, *11*, 60-72.

Millsap, R. E. (2001). When trivial constraints are not trivial: The choice of uniqueness constraints in confirmatory factor analysis. *Structural Equation Modeling*, *8*, 1-17.

Millsap, R.E. (2005). Four unresolved problems in studies of factorial invariance. In A. Maydeu-Olivares & J.J. McArdle (Eds.) *Contemporary Psychometrics* (pp. 153-172). Mahwah, NJ: Lawrence Erlbaum Associates.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory, *Journal of Applied Psychology*, *87*, 517-529.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches to exploring measurement invariance. *Psychological Bulletin*, *114*, 552-566.

Rensvold, R. B., & Cheung, G. W. (2001). Testing for metric invariance using structural equation models: Solving the standardization problem. In C. A. Schriesheim & L. L. Neider (Eds.), *Research in management: Equivalence in measurement* (pp. 25-50). Greenwich, CT: Information Age Publishing.

SAS Institute (2004) *SAS version 9.1.3*. Cary, NC: SAS Institute, Inc.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, D.C.: American Psychological Association.

Vandenberg, R. J. (2002). Toward a further understanding of and improvement in measurement invariance methods and procedures. *Organizational Research Methods*, *5*, 139-158.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-69.

Wilson, K. L. (1981). On population comparisons using factor indexes or latent variables. *Social ScienceResearch*, *10*, 301-313.

Yoon, M., & Millsap, R. E. (2007). Detecting violations of factorial invariance using data-based specification searches: A monte carlo study. *Structural Equation Modeling*, *14*, 435-463.

Zumbo, B. D. (2003). Does Item-Level DIF Manifest Itself in Scale-Level Analyses?: Implications for Translating Language Tests. *Language Testing*, *20*, 136-147.